Life-Sized Audiovisual Spatial Social Scenes with Multiple Characters: MARC & SMART-I2

Matthieu Courgeon, Marc Rébillat, Brian Katz, Céline Clavel, Jean-Claude Martin

LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

ABSTRACT

With the increasing use of virtual characters in virtual and mixed reality settings, the coordination of realism in audiovisual rendering and expressive virtual characters becomes a key issue. In this paper we introduce a new system combining two systems for tackling the issue of realism and high quality in audiovisual rendering and life-sized expressive characters. The goal of the resulting SMART-MARC platform is to investigate the impact of realism on multiple levels: spatial audiovisual rendering of a scene, appearance and expressive behaviors of virtual characters. Potential interactive applications include mediated communication in virtual worlds, therapy, game, arts and e-learning. Future experimental studies will focus on 3D audio/visual coherence, social perception and ecologically valid interaction scenes.

1 INTRODUCTION

With the increasing use of virtual characters in virtual and mixed reality settings, the coordination of realism in audiovisual rendering and expressive virtual characters becomes a key issue.

The advancement of immersive environments has separately produced systems with improved quality for 3D stereoscopic graphical rendering [1] and also for spatialized audio rendering [2-4]. Despite these advances, few combined modality systems of high quality have been developed (see [5, 6] for example). This can be mainly attributed to the different and stringent technical requirements that are needed to render each modality with a high degree of quality.

Similarly, virtual characters are now able to display facial expressions of emotions and postural expressions of interpersonal attitudes. Experiments involving virtual characters in immersive environments have highlighted *presence* as a key feature related to a user's perception of these virtual characters. The impact of realism of expressive virtual characters are also investigated in terms of appearance or behaviors. Yet, few platforms enable one to investigate both the realism and the interactivity of expressive characters at the same time.

In this paper we introduce a new system combining two systems for tackling the issue of realism and high quality in audiovisual rendering and expressive characters. SMART-I² (Spatial Multi-user Audio-visual Real-Time Interactive Interface) is a high quality 3D audio-visual immersive interactive rendering system [7, 8]. MARC (Multimodal Affective and Reactive Character) [9-11] is a platform for real-time affective interaction with multiple characters. The goal of the resulting SMART-MARC platform is to investigate the impact of realism on multiple levels: spatial audiovisual rendering of a scene, appearance and expressive behaviors of virtual characters. Potential interactive applications include mediated communication in virtual worlds, therapy, game, arts and e-learning.

Section 2 describes related work. Section 3 details the integration of the SMART-I² and MARC. Section 4 explains the future experimental studies and applications that will be possible with this integrated platform.

2 RELATED WORK

2.1 Affective Virtual Characters

Embodied conversational agents are computer-generated characters that aim to demonstrate the same properties as humans in face-to-face conversation, including the ability to produce and respond to verbal and nonverbal communication [12]. They need to be able to express emotion in multiple modalities (e.g. face, gesture and voice). Expression of basic emotions using these modalities has been extensively studied for the past thirty years, progressively increasing realism and credibility of virtual agents [13].

More recently, technology has evolved to a point where multiple characters can be displayed and rendered simultaneously. Simulating and evaluating social interactions with multiple characters is now a key challenge for such systems leading to recent prototypes and experimental studies.

Following the social inhibition and social facilitation theory, the mere presence of a virtual character was observed to have an impact on the achievement of a task by users; the presence of the virtual character improving user performance on simple tasks and decreasing performance for complex tasks [14].

Several studies have explored the coordination requirements in a small group of avatars using desktop PC configurations. The importance of simulating territorial behavior in a small group of conversing avatars was observed by [15]. They designed the "Populous platform" which simulates spatial behavior of small groups of virtual characters. Each character takes into account interpersonal distance according to social rules. Bodily and gaze attitudes of dominance and affiliation were also simulated in conversing avatars [16]. Dialog turn-taking management was simulated using up to twelve virtual characters. The Ymir Turn Taking Model is a cognitive model of multimodal realtime turntaking that has been tested in numerous two-party dialogue systems. It was recently extended to manage multiple character scenes in real-time [17].

Several studies displayed virtual characters in immersive and mixed reality set-up. The concept of *presence* is important in such a setup. Presence refers to experiencing the computer-generated environment rather than the actual physical local one [18], experiencing the scene, and not the technology. Bailenson *et al.* extended this concept to *social presence*, occurring when virtual characters are displayed in a virtual environment [19]. The impact of a mixed reality display configuration on user behavior with a virtual human was also investigated. The virtual character was judged as more engaging, empathetic, pleasant, and natural while displayed by mixed reality on a 20 inch plasma TV, than rendered on a standard desktop screen.

Despite these findings, few systems enable users to interact with multiple characters in an immersive set-up. This paper presents a new platform, displaying semi-realistic virtual characters using immersive audiovisual spatial rendering where modality coherence is a key design goal.

2.2 Audiovisual spatial rendering

In any audio-visual application, the sensation of immersion [20] and the intelligibility of the scenes [21], which are closely related to presence, depend highly on the spatial quality provided by both the audio and the visual renderings. It is thus important to find ways to fulfill the requirements for both audio and visual modalities, achieving a device which provides a perceptually coherent spatial audio-visual rendering over a large rendering area. Such a rendering device thus realizes a seamless transition between the real and virtual worlds. As so, a device which ideally renders a spatial audio-visual scene can be thought of as a “large open window through which the users experience the virtual world” [22].

The key concept is therefore to create this virtual audio-visual window through which a plausible virtual world is perceived. All the spatial properties of the audio-visual scene must then be accurately conveyed to the user for any position within the rendering area, including angular and distance perception of objects which should remain accurate throughout. The audio-visual window therefore ensures that both static and also dynamic localization cues, such as the motion parallax effect, are preserved for both auditory and visual virtual objects

3 SYSTEM

3.1 SMART-I²

The SMART-I² system [7, 8] is able to realize a high degree of 3D audio-visual integration with almost few compromises on either the audio or the graphics rendering quality. Moreover users can progress and move about within a large rendering area while still perceiving coherent audio-visual localization cues coming from the different virtual characters and their surrounding environment.

3.1.1 Sound rendering using Wave Field Synthesis

Wave Field Synthesis (WFS) is a spatialized sound rendering technology which was first really developed at Delft University [4]. It is an audio implementation of Huygen’s principle which states that: *“Every sound field emerging from one primary sound source can be reproduced by summing contributions of an infinite and continuous distribution of secondary sound sources”*. At the theoretical level, WFS allows one to synthesize a sound source at any given position. Implementations of WFS are simplified versions of this principle, typically using a linear array of equally spaced loudspeakers.

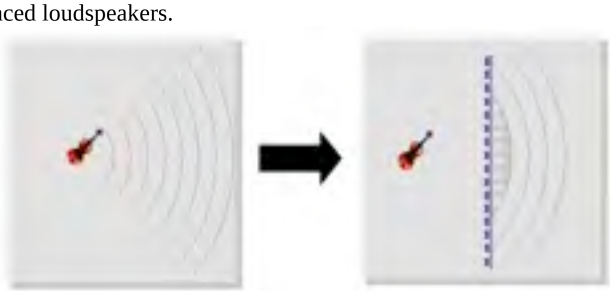

Figure 1: Illustration of sound rendering using WFS

Figure 1 illustrates the principle of WFS. The violin on the left part is the primary source producing the target natural sound field. The linear array of secondary sound sources on the right produces, through summation of the contributions of each loudspeaker driven appropriately, a synthesized sound field equivalent to the original target field. The sound field of the virtual violin is synthesized, perceived by users in the reproduction area as emanating from the precise spatial location of the violin. Additional sound sources may be simultaneously synthesized through simple linear superposition.

Using this physical basis, different types of fundamental sound sources, or sound fields, can be synthesized (Figure 2). Plane waves represent sound objects situated far away from the immersion area and are perceived as coming from a constant angle, independent of listener position. Point sources represent sound objects near the immersion area. Point sources can be synthesized at positions behind or in front of the loudspeaker array. Such focused sources (i.e. point sources in front of the array) are perceived as being physically present in the immersion area.

Figure 2: Synthesis of different wave types. *Left*: Point source behind the loudspeaker array. *Center*: Plane wave. *Right*: Point in front of the loudspeaker array

Sound rendering using WFS, due to its approach in physically recreating the entire field within a spatially large area, is not limited to a single user at a single location. The sound perspective, including parallax, is correct for every user in the immersion area, without the need for a tracking device.

As previously stated, practical WFS implementations are limited to a linear array and hence reproduction is optimized for the horizontal plane. Auditory perception is the most precise and more stable in the horizontal plane and therefore a more pertinent choice for array orientation. With this restriction, which reduces the required calculation power and audio hardware, the digital audio processing can be done with a latency of less than 5 ms. This is more than sufficient for real-time AV applications.

3.1.2 Visual rendering with tracked passive stereoscopy

To produce a 3D visual rendering, each eye of the user must see the scene from a slightly different point of view. One means of realizing this is to use light polarization properties to independently address each eye of the user. The user wears special polarized glasses for visual cross-talk cancellation. The graphic rendering should also be adapted to the user’s head position in order to maintain the correct point of view. Using this approach, the 3D visual rendering is coherent regardless of the user’s position in the viewing area. This technique is referred as tracked passive stereoscopy (TPS).

3.1.3 Integration using Large Multi-actuator Panels

The seamless integration of the two different technologies (WFS and TPS) is achieved through an innovative use of multi-actuator panels (MAPs) [24]. MAPs are stiff lightweight panels with multiple electro-mechanical exciters attached to the backside. For this project, a novel large dimension MAP (LaMAP) has been designed, (~5 m² with a 4/3 ratio) in order to provide sufficient surface area and size to be used as a projection screen in an immersive scenario. To accommodate polarized light projection, the front face of the panel has been treated with metallic paint designed to preserve light polarization. Due to the nature of the LaMAP design, screen displacements caused by acoustic vibrations are very small and are not visible with regards to the 3D video projection on the surface of the panel.

Such a structure allows one to efficiently integrate a 3D visual rendering technology (TPS) and a spatialized sound rendering technology (WFS). Currently, the SMART-I² is made of two large MAPs of 2.6 m x 2 m which form a corner of stereoscopic screens and a 24 loudspeakers array (12 ch / screen). With this configuration, users can move within an immersion area of approximately 2.5 m x 2.5 m, as can be seen in Figure 3.

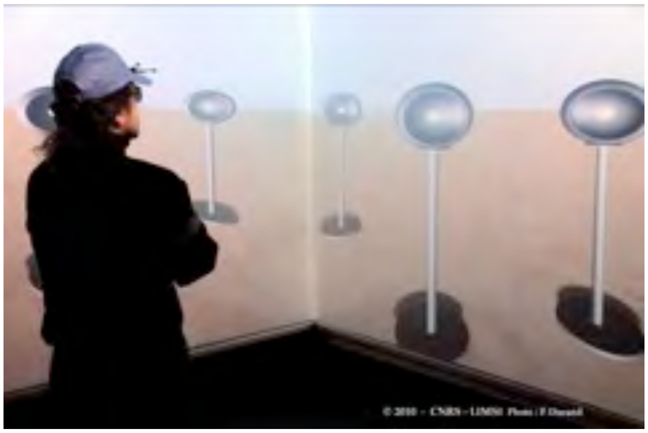

Figure 3: An example of a simple audio-visual scene provided by the SMART-I²

3.2 MARC

MARC (Multimodal Affective and Reactive Characters) is a framework for real-time affective interaction with multiple characters. Simulation of affective behaviors is achieved using implementations of several emotional models: categorical models of emotion [9], P.A.D. [23] dimensional model [11] and CPM [24] cognitive evaluation models [10].

In addition to our research, MARC was applied to several domains, such as arts, or clinical application on autistic children [25].

MARC features three main modules: facial expressions edition, body gesture edition, and real-time interactive rendering. MARC's real-time rendering relies on GPU programming (OpenGL/GLSL) to render detailed models and realistic skin lighting (shadow casting, BSSRDF: simulation of light diffusion though skin). Our facial animation system extends the MPEG-4 model [26] by adding wrinkle management (Figure 4). Key expressions are predefined as a set of keypoint displacements. The real-time rendering system performs animation by blending several key expressions. This technique also enables one to perform Action Unit based animation. Action Unit expressions are then considered as MPEG-4 key expressions.

Figure 4. MARC facial edition and rendering

3.2.1 Manual Emotional Control

Several techniques where explored to control the character's expressivity. The recognition software FaceReader is used to decode emotions expressed by a user's face. This signal is modulated by expressive profiles (High expressivity, Low expressivity, Negative, or Positive) and reproduced on the virtual character's face. (Figure 5 Top). Another approach is to use 3D devices, such as 3D mice or joystick, to control the emotional state of the agent, represented in a 3D space (Pleasure, Arousal, Dominance). Then, we apply the modulation using expressive profiles, and the resulting emotion is display on the virtual character's face (Figure 5 Bottom).

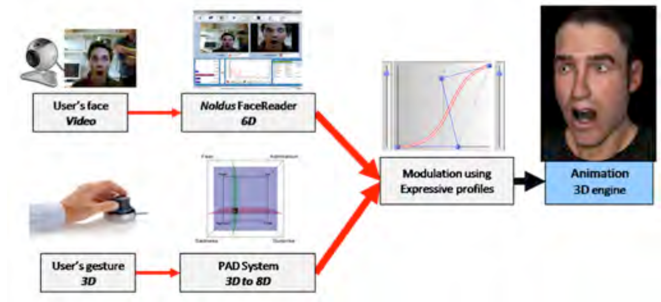

Figure 5. Interaction for emotional control

3.2.2 Automatic emotional control

We developed emotional reaction modules based on cognitive evaluation theories [24]. The virtual character appraises the game's events and situation according to several cognitive criteria (novelty, expectedness, pleasantness, goal conductivity, coping ability, external/internal causation, internal norms, and external norms). Figure 6 presents the interactive architecture. These modules were evaluated during a real-time interactive game with the virtual character in a non-immersive environment.

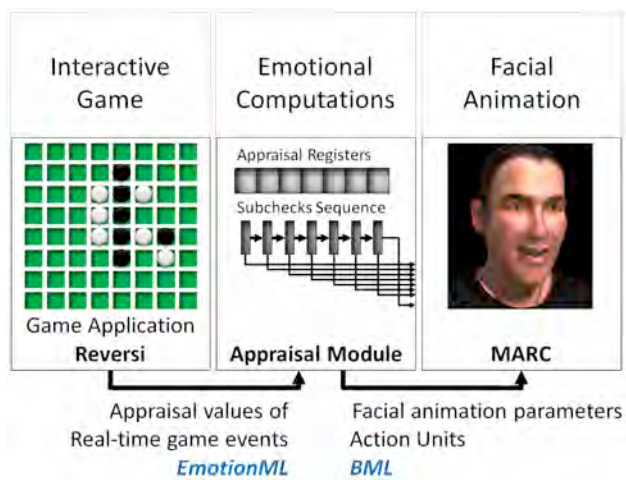

Figure 6. Architecture of the autonomous emotional reactions during a real-time game

3.2.3 Posture edition and animation

Body animation is based on skeleton rigging (Figure 7). The body gesture editor enables to edit sequences of key poses. MARC features two ways to create posture animations: 1) the animation designer directly manipulates the bone positions and rotations to edit each posture composing the animation, 2) the sequence of postures is generated using a motion capture file (BVH format).

3.2.4 Real-time interaction

MARC real-time rendering is controlled by external applications using a standard protocol, (Behavior Markup Language [27]). It uses predefined key animations (both face and body) to create a dynamic interactive animation.

The MARC framework enables one to integrate virtual characters in various 3D environments. These environments can be edited using 3D software such as blender, and exported in X3D standard format. MARC reconstructs the scene using an environment file describing which objects are in the scene and where they are. Environment sounds can also be added, as WAV files with a corresponding spatial position (Figures 8 and 9).

MARC currently provides three models, two adult male, and one adult female (Figure 10).

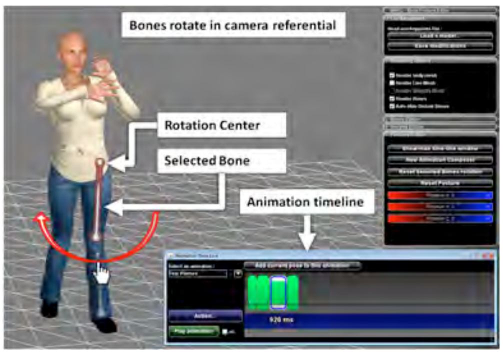

Figure 7. MARC's body posture editor

Figure 8. Environment: "Hall"

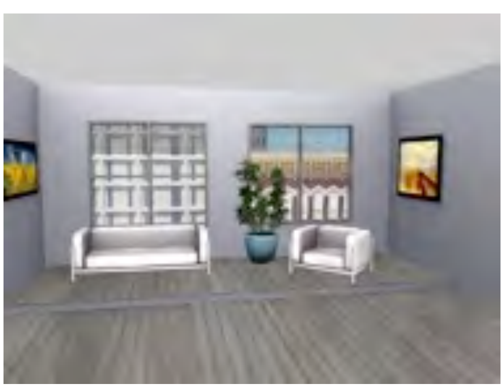

Figure 9. Environment : "Apartment"

Figure 10: The Three MARC's characters.

3.3 SMART-MARC: Integration of SMART-I² and MARC

The MARC and the SMART-I² systems were integrated in order to enable high quality spatial rendering of virtual social scenes. Specific network protocols where defined to exchange information such as tracking information, and sound localizations in the 3D virtual space. An overview of the architecture of the integrated system is shown in Figure 11.

Max/MSP patches compute information relative to sound spatialization, handle individual source audio through files or synthesis, and control the WFS engine. The WFS engine performs the real time signal processing necessary to spatialized the received audio sources which are rendered over the 24 loudspeaker array. MARC computes information relative to the visual modality, such as frustum deformation (adapting screen aspect to user location).

Due to this architecture, synchronization between Max/MSP and MARC is a crucial aspect for audio-visual congruence. To address this issue, we designed synchronization protocols. Loading and pre-computations are performed independently on both sides, and then "bang" signals trigger synchronized audio-visual animation.

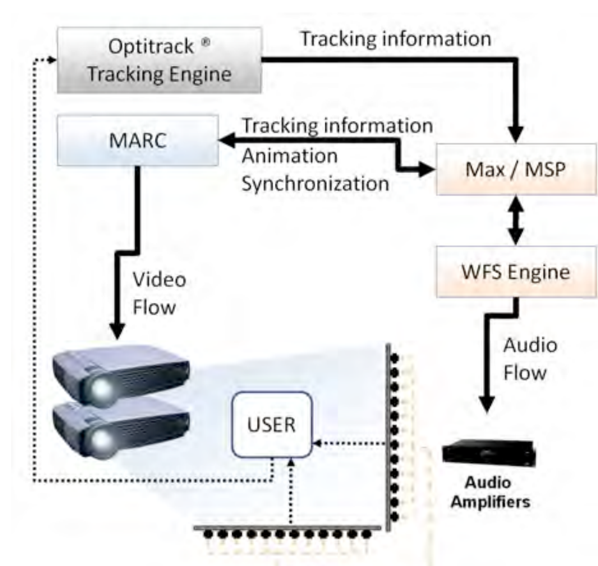

Figure 11. Integration of MARC and SMART-I2

Figure 12 shows an example of a user interacting with MARC within the SMART-I² virtual environment.

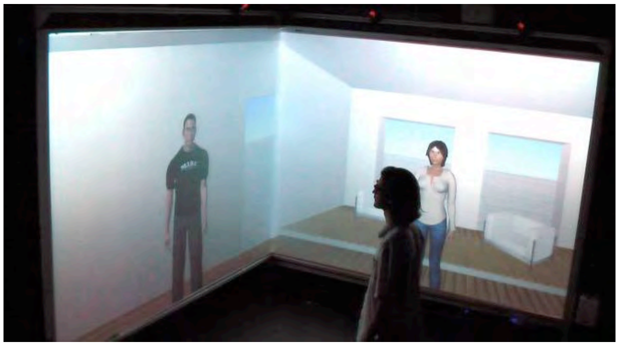

Figure 12. The SMART-I² environment rendering two MARC characters

4 CONCLUSIONS AND FUTURE DIRECTIONS

We introduced a new platform for high quality spatial rendering of virtual scenes featuring multiple virtual characters.

The SMART-I2+MARC system is being used for experimental studies testing the influence of coherent audiovisual rendering and immersion on the perception the user has of multimodal expressions of emotions displayed by virtual characters.

5 ACKNOWLEDGEMENTS

The SMART-I2 system is a joint project between LIMSI-CNRS and *sonic emotion*, with funding received through the Action Initiative program for immerging research at LIMSI-CNRS.

REFERENCES

[1] Froehlich, B., Blach, R., and Stefani, O.: ‘Implementing Multi-Viewer Stereo Displays’. Proc. WSCG Conference Proceedings, Plzen, Czech Republic, 2005 pp.

[2] Gerzon, M.A.: ‘Periphony: With-Height Sound Reproduction’, Journal of the Audio Engineering Society, 1973, 21, (1), pp. 2-10

[3] Pulkki, V.: ‘Virtual sound source positioning using vector base amplitude panning’, Journal of the Audio Engineering Society, 1997, 45, (6), pp. 456-466

[4] Berkhout, A.J., de Vries, D., and Vogel, P.: ‘Acoustic Control By Wave Field Synthesis’, Journal of Acoustical Society of America, 1993, 93

[5] Kuhlen, T., Assenmacher, I., and Lentz, T.: ‘A true spatial sound system for CAVE-like displays using four loudspeakers ’. Proc. International conference on Virtual reality 2007 pp. 270-279

[6] Springer, J.P., Sladeczek, C., Scheffler, M., Hochstrate, J., Melchior, F., and Froehlich, B.: ‘Combining Wave Field Synthesis and Multi-Viewer Stereo Displays’. Proc. Virtual Reality, Washington, USA 2006 pp. 237-240

[7] Rébillat, M., Corteel, E., and Katz, B.: ‘SMART-I² “Spatial Multi-user Audio-visual Real-Time Interactive Interface’. Proc. Convention of the Audio Engineering Society 2008

[8] Rébillat, M., Corteel, E., and Katz, B.F.G.: ‘SMART-I²: Spatial Multi-users Audio-visual Real Time Interactive Interface, a broadcast application context’. Proc. 3DTV Conference, Potsdam, Germany 2009

[9] Courgeon, M., Buisine, S., and Martin, J.-C.: ‘Impact of Expressive Wrinkles on Perception of a Virtual Character's Facial Expressions of Emotions’. Proc. 9th International conference on Intelligent Virtual Agents (IVA09), Amsterdam, The Netherlands 2009 pp. 201-214

[10] Courgeon, M., Clavel, C., and Martin, J.-C.: ‘Appraising Emotional Events during a Real-time Interactive Game’. Proc. ICMI09 International workshop on Affective-aware Virtual Agents and Social Robots (AFFINE09), Boston, MA, 2009 pp.

[11] Courgeon, M., Martin, J.-C., and Jacquemin, C.: ‘User's Gestural Exploration Of Different Virtual Agents' Expressive Profiles’. Proc. 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS'2008), Estoril, Portugal 2008 pp. 1237-1240

[12] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E.F.: ‘Embodied Conversational Agents’ (2000)

[13] Pasquariello, S., and Pelachaud, C.: ‘Greta: A Simple Facial Animation Engine’, 6th Online world conf. on soft computing in industrial applications, 2001

[14] Hayes, A., Ulinski, A., and Hodges, L.: ‘That Avatar Is Looking at Me! Social Inhibition in Virtual Worlds’. Proc. Intelligent Virtual Agents 2010 pp. 454-467

[15] Pedica, C., and Vilhjálmsson, H.H.: ‘Spontaneous Avatar Behavior for Human Territoriality ’. Proc. Intelligent Virtual Agents 2009 pp. 344-357

[16] Gillies, M., and Ballin, D.: ‘A model of Interpersonal Attitude and posture generation’. Proc. Intelligent Virtual Agents, 2003 pp. 88-92

[17] Thórisson, K., Gislason, O., Jonsdottir, G., and Thorisson, H.: ‘A Multiparty Multimodal Architecture for Realtime Turntaking’: ‘Intelligent Virtual Agents’ (Springer Berlin / Heidelberg, 2010), pp. 350-356

[18] Slater, M.: ‘Measuring presence: A response to the Witmer and Singer presence questionnaire’, Presence: Teleoperators and Virtual Environments, 1999, 12, (5), pp. 456-480

[19] Bailenson, J.N., Blascovich, J., Beall, A.C., and Loomis, J.M.: ‘Interpersonal Distance in Immersive Virtual Environments’, Personality and Social Psychology Bulletin 2002, 29, pp. 1-15

[20] Blauert, J.: ‘Communication Acoustics’, Signals and CommunicationTechnology, 2005

[21] Bruijn, W.P.J.d.: ‘Application of wave field synthesis in videoconferencing’, Delft University of Technology, 2004

[22] Brooks, F.P.: ‘ What's real about virtual reality?’, IEEE Computer Graphics And Applications, 1999, 19, pp. 16-27

[23] Russell, J.A., and Mehrabian, A.: ‘Evidence for a three-factor theory of emotions’, Journal of Research on Personality, 1977, 11, (3), pp. 273–294

[24] Scherer, K.R.: ‘Appraisal considered as a process of multilevel sequential checking’, in Klaus R. Scherer, A.S., Tom Johnstone, Eds (Ed.): ‘Appraisal processes in emotion: theory, methods, research’ (Oxford University Press US, 2001, 2001), pp. 92-120

[25] Grynszpan, O., Nadel, J., Constant, J., Barillier, F.L., Carbonell, N., Simonin, J., Martin, J.-C., and Courgeon, M.: ‘A new virtual environment paradigm for high functioning autism intended to help attentional disengagement in a social context’, Special Issue of the internationnal "Journal Of Physical Therapy Education" 2010

[26] Pandzic, I., and Forchheimer, R.: ‘MPEG-4 Facial Animation: The Standard, Implementation and Applications’ (John Wiley \& Sons, Inc., 2003)

[27] Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafa, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., Welbergen, H.v., and Werf, R.J.v.d.: 'The behavior markup language: recent developments and challenges'. Proc. 7th International Conference on Intelligent Virtual Agents, Paris, France, 2007 pp. 90–111